APPENDIX E-4
WEB SERVER TRANSACTION LOG ANALYSIS RESULTS

APPENDIX E-4
WEB SERVER TRANSACTION LOG ANALYSIS RESULTS

1.0. INTRODUCTION

Use of the Internet in general, and the Web in particular, continues to increase dramatically. Indeed, as of January 1997, there are 16,146,000 Internet-based hosts and 828,000 domains (Network Wizards, 1997). This is nearly double the number of hosts and triple the number of domains as compared to January 1996 (Network Wizards, 1997).

Along with for-profit and non-profit organizations, Federal government agencies are increasing their use and provision of electronic networked services. Moreover, agencies continue to devote additional resources to the development and maintenance of Web-based services. Several critical Web service-related questions face the providers of such services:

- What is the server's traffic and overall ability and necessary resources to meet the demands of that traffic?
- What is the server's user community, as identified by the
  - accessing host IP address?
  - type of browser and operating system?
- What did users do while interacting with the server?
- From where did a user access and at what point leave the server?
- What problems did users encounter during their server sessions?

One means through which agencies can begin to answer these questions is through the analysis of Web-server generated log files. This appendix presents an overview of Web statistics, the process of analyzing and interpreting Web log files, and methodological details and findings of the research team's analysis of 14 days of Web server log files from the Environmental Protection Agency (EPA).

2.0. WEB SERVER LOG ANALYSIS: STEPS AND TOOLS

Log analysis is essentially a three step process that involves planning, data analysis, and interpretation activities. In particular, there is a need to:

- ***Determine what types of information server administrators and decision makers need***. Log analysis is one means through which to determine whether Web-based services are meeting their intended missions or objectives. As such, server administrators and decision makers need to know what types of information are wanted prior to the analysis of Web server log files so as to ensure the collection of data that will assist in assessing mission or goal attainment.
- ***Develop a program that can parse through, manipulate, and present value-added information from the log files.*** Server administrators have the option of writing their own programs, downloading free software, or purchasing one of many off-the-shelf analysis products to do this. A listing of numerous Web analysis programs is available from webreference.com <http://www.webreference.com/usage.html>. Although continually increasing in their analysis capabilities, most programs tend to only parse through specific variables, leaving many important pieces of information untouched.
- ***Analyze the information generated from the program***. This is not as straight-forward a process as one might think. For example, most log analysis software programs analyze the number of "hits" -- not accesses -- a server receives. In this case, the hit count reflects the number of items (e.g., images) downloaded when a user accesses a particular page. So, if a site has a corporate logo image file on every page, that image will more than likely be the most frequently downloaded -- "hit" -- item on the site. Analysis information such as that is relatively useless in determining the site's actual usage.

The program selection and analysis processes complement each other. Depending on the log analysis software used, server administrators are limited to certain types of log analysis. Based on those analysis limitations, server administrators need to know the meaning of the log analysis output (e.g., whether the statistics represent "hits" or "accesses.").

There are several Web server analysis software packages available on the market today, both free and for fees. Readers interested in reviewing sample Web analysis software should refer to the following sites: <http://www.statslab.cam.ac.uk/~sret1/analog/>; <http://www.mkstats.com>; <www.ics.uci.edu/pub/websoft/wwwstat/>; and, <www.boutell.com/ wusage/intro2.html>. In addition, readers should review material found in Stout (1997). Readers may find the above listed log analysis software of use, depending on their analysis needs and requirements.

As discussed in the Methods section below, the study team assessed the currently available log analysis software packages and found them inadequate to perform the various analysis of the EPA log files of interest to the GILS evaluation study. The study team, therefore, developed its own PERL-based analysis scripts to analyze the EPA log files.

2.1. Background on Web Server Log Files

Web servers automatically generate and dynamically update four usage log files. These four log files and types of information each captures are as follows:

- Access Log (e.g., hits);
- Agent Log (e.g., browser, operating system);
- Error Log (e.g., download aborts); and
- Referer Log (e.g., referring links).

The log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a given a web site (for additional background information, refer to Rubin, 1996; Noonan, 1996; Novak and Hoffman, 1996).

In determining the amount of traffic a site receives during a specified period of time, it is important to understand what, exactly, the log files are counting and tracking. In particular, there is a critical distinction between a ***hit*** and an ***access***, wherein:

- A ***hit*** is any file from a web site that a user downloads. A hit can be a text document, image, movie, or a sound file. If a user downloads a web page that has 6 images on it, then that user "hit" the web site seven times (6 images +1 text page).

- An ***access***, or sometimes called a page hit, is an entire page downloaded by a user regardless of the number of images, sounds, or movies. If a user downloads a web page that has 6 images on it, then that user just accessed one page of the web site.

This distinction is noteworthy. Most web analysis software counts the number of ***hits*** a server receives, rather than the number of ***accesses***.

2.1.1. Access Log

The Access Log provides the greatest amount of server data, including the date, time, IP address, and user action (e.g., document/image/sound/movie download). The following is an example line of text from an Access Log:

smx-ca8-50.ix.netcom.com - - [30/Sep/1996:02:57:07 -0400] "GET/Proj/main.html

It is possible to analyze the following variables in the Access Log:

- ***Domain name or Internet Protocol (IP) number***. In the above example we know that the user's computer had the following domain name: smx-ca8-50.ix.netcom.com.

- ***Date and Time***. In the above example we know that the user accessed a page on September 30, 1996 at 2:57 AM and 07 seconds. By default the time is based on a twenty four hour clock.

- ***Item accessed.*** The word item can mean an image, movie, sound, or html file. The above example shows that main.html was the item accessed. It is also important to note that the full path name (from document root) is given to avoid confusion, (e.g., there may be more then one main.html on a server).

It is possible to generate the following data from these variables,:

- The percentage of users accessing the site from a specific domain type (e.g., .com, .edu, .net, .mil, .gov). This can be analyzed further by ***hits*** versus ***accesses***.

- The number of hits the server is getting from various IP groups. Such data can inform server administrators as to the primary clients of their servers.

- The number of unique IP addresses accessing the site. While not a measure of unique users, this can provide server administrators with some indication of the number of users by stripping repeat IP addresses from the log data. This data is an important indicator of the breadth of penetration of a server.

- The number of accesses/hits the server receives during specific hours and days of the week. These statistics can be useful to server administrators who need to know the optimal time/day to perform server maintenance and/or upgrades.

- The path -- known as “threading” -- a user takes through a site. Knowing this allows server administrators to determine the average length of a user’s session, specific location duration (e.g., average time on a page), average download times, and how the user navigated through the site (e.g., entrance and exit points).

The data from the Access Log provides a broad view of a Web server’s use and users (as indicated by IP addresses). Such analysis enables server administrators and decision makers to characterize their server’s audience and usage patterns.

2.1.2. Agent Log

The Agent Log provides data on a user’s browser, browser version, and operating system. This is significant information, as the type of browser and operating system determine what a user is able to access on a site (e.g., Java, forms). Below is a sample Agent Log entry:

Mozilla/3.0 (Win95; I)

Analysis of the Agent Log enables server administrators to determine the (see Figures 6-8):

- ***Browser***. The type of browser used to access a web site. There are several different Web browsers on the market today (e.g., Netscape, Microsoft Internet Explorer, LYNX, Mosaic), each of which have different viewing capabilities.

- ***Browser version***. The version of a browser used. Not all browsers can view all components of a Web site. For example, Netscape version 1.0 cannot view forms-based data.

- ***Operating system***. The type of computer and operating system users have A Web site can look different to users depending on their computer platform (e.g.,Windows, Win95, Macintosh, PowerPC, SunOS).

These data are essential for the design and development of Web sites. Without such information, server administrators could design sites that require viewing capabilities that a vast majority of the site’s users do not possess. At best, this leads to wasted effort by the server administrator. At worst, this can lead to improperly displayed Web content, thus effectively rendering the site useless to the user.

2.1.3. Error Log

The average Web user will receive an "Error 404 File Not Found" message several times a day. When a user encounters this message, an entry is made in the Error Log. Below is a sample Error Log entry:

[Sun Nov 3 23:57:00 1996] httpd: send aborted for pm02_23.ct.net,
URL:/OWOW/images/new/owpool.gif

The Error Log contains the following data for analysis:

- ***Error 404***. The Error Log tells a server administrator the time, domain name of the user, and page on which a user received the error. These error messages are critical to Web server administration activities, as they inform server administrators of problematic and erroneous links on their servers.

- ***Stopped transmission***. This informs a server administrator of a user-interrupted transfer. For example, a user clicking on the "stop" button would generate a "stopped transmission" error message. The Error Log tells a server administrator the time, domain name, and page that a user was on when the transmission was stopped (as in the above sample Error Log entry). This information is useful as it can indicate patterns with large files such as image, movie, and other files that users consistently stop downloading.

The analysis of Error Log data can provide important server information such as missing files, erroneous links, and aborted downloads. This information can enable server administrators to modify and correct server content, thus decreasing the number of errors users encounter while navigating a site.

2.1.4. Referer Log

The Referer Log indicates what other sites on the Web link to a particular server. Each link made to a site generates a Referal Log entry, a sample of which is below:

http://www.altavista.digital.com/cgibin/query?pg=q&what=web&fmt=.&q=SIC+CODE ->
/xxx/html/rcris/rcr_sic_code.html

In this particular example, the referer was AltaVista, indicating that the user entered the Web site after performing a search using the AltaVista search facility.

The Referer Log entry provides the following data:

- ***Referral***. If a user is on a site (e.g., ericir.syr.edu), and clicks on a link to another site (e.g., www.sun.com), then www.sun.com will receive an entry in their Referer Log. The log will show that the user came to the sun site (www.sun.com) via ericir.syr.edu (the referral).

Such referral data is critical to alleviating missing link (Error 404) data. For example, when the URL of a page within www.sun.com changes, the server administrator of www.sun.com could notify all referrals (e.g., ericir.syr.edu) of the change. This can alleviate future "Error 404 - File Not Found" messages.

Through the analysis of the four log files, Web service providers can begin the process of assessing and evaluating their networked information services. Current Web usage statistics generally center on the analysis of the Access Log, thus limiting the ability of Web-mounted service extensiveness measures. There are, however, means to analyze the Agent, Error, and Referer log files. Such techniques can provide important additional insight into the use of Web-based services by users.

3.0. METHODOLOGY

As part of the evaluation study of U.S. Implementation of GILS, the authors selected one Federal agency's Web server from which to collect log files. The authors performed analysis on a sample of the log files to:

- Determine the overall Web site's traffic, including the

  - origin of users
  - portions of the site that are accessed
  - number of document downloads (both hits and accesses);

- Determine the use of the Web site GILS traffic, including the

  - origin of users

  - portions of the site that are accessed
  - number of document downloads (both hits and accesses);

- Experiment with developing new log analysis techniques that go beyond domain, hit, and browser counts;

- Assist Federal agencies that operate Web-based GILS servers to develop, implement, and maintain on-going log file analysis; and

- Inform Federal agencies that operate Web-based GILS servers of the utility in analyzing and interpreting log file data in on-going assessments of their GILS implementations.

Such an evaluation enables the maintainers, policy makers, and stakeholders of the Web site to determine a site's use as one component of an overall networked information resource.

The log files were collected daily between February 2, 1997 and February 15, 1997. The four log files ranged in size from 8 megabytes to 26 megabytes each per day. In all, approximately 560 megabytes of log file data were collected. The resulting output, Web log file analysis PERL scripts, and log files together consumed approximately 1 gigabit of storage. The analysis of the EPA log files was performed on a Pentium 150 MHZ computer with 32 MB of RAM, and the analysis of each of the four daily log files took approximately 40 minutes.

3.1. Choosing Web Analysis Software

The authors reviewed multiple Web analysis software packages along the following criteria:

- Ability to provide global and directory specific Web server analysis;

- Ability to distinguish between hits and accesses;

- Ability to determine user-specific actions (e.g., navigation) through a Web site session; and

- Ability to distinguish between unique and total referals.

Most existing log analysis software could perform one or more of the above functions. None, however, met all the analysis criteria for the GILS evaluation project. As such, the authors worked with a study team at the School of Information Studies, Syracuse University, to develop PERL-based Web analysis scripts that would provide all the required analysis capabilities. Readers, therefore, will find that currently available Web log file analysis software cannot perform some of the analysis techniques presented in this appendix.

3.2. Developing the PERL Scripts

The development and pre-testing of the PERL scripts required considerable effort. The Syracuse University script development team required the equivalent of 240 man-hours developing the scripts. An additional 100 man-hours were required to pre-test the scripts using several different log files from different servers, including a test data set from the Federal agency HTTP GILS server. Running the scripts on the 14 day period of EPA log files and outputting the analysis into a usable format required an additional 100 man-hours. In total, therefore, the PERL script development process consumed approximately 420 man-hours.

To ensure valid and reliable results, script file results were compared to results generated by other log analysis software, where possible. When errors in script files were found, corrections were made and the files re-tested.

4.0. FINDINGS

The study team analyzed each of the four log files on a daily basis. Analysis of the files is presented both in aggregate and individual day format where possible. To simplify the presentation of the data, the findings are presented by log file type. Readers should note that these findings do not include such commonly available analysis as hits by time of day or day or week. Rather, this appendix presents findings from the use of the developed PERL scripts intended to provide new and previously unavailable forms of log analysis.

4.1. Access Log

The EPA Web server generates considerable traffic on any given day (see Tables 1 and 2). On average, the EPA server receives approximately 80,000 daily accesses that generate over 213,000 daily hits (see Tables 1 and 2). In all, the EPA server received over 564,000 accesses per week generating over 1,496,000 hits per week. As Tables 1 and 2 demonstrate, the EPA server is most used during the middle part of the week.

On average, the GILS component of the EPA server (As measured by use of the Earth100 directory), the daily percentage of GILS accesses ranges from .45% to .93%, with a weekly average of .52% and .61%, respectively (see Tables 3 and 4). These GILS accesses account for .20% to .44% of all EPA server hits. As with the EPA server in general, the GILS portion of the server is most heavily used during the middle of the week.

It is important to note three factors when considering the average GILS usage patterns as depicted in Tables 3 and 4:

- The tables do not include Z39.50 accesses to GILS records, thus do not necessarily reflect the total usage of the EPA GILS database; and

- The tables do not compare, nor did the study collect such data, the traffic the GILS component of the EPA server to other EPA server components. The EPA server has a significant number of subdirectories that would require traffic analysis to gain a more accurate sense of the GILS directory traffic in relation to other server directories.

The overall use of the EPA GILS records is underreported in Tables 3 and 4 without such data.

It is interesting to note that the EPA server in general, and the GILS portion in particular, both receive a fairly consistent percentage of traffic (as measured by accesses and hits) from within the United States and from foreign countries (see Tables 5 and 6). The daily average for United States-generated EPA accesses ranges from 72.48% to 74.19%, while the daily average accesses from foreign countries ranges from 25.81% to 27.52% (see Tables 5 and 6). The daily average for United States-generated GILS accesses ranges from 28.39% to 35.06%, while the daily average accesses from foreign countries ranges from 28.39% to 35.06% (see Tables 5 and 6). From the limited data set, it is not possible to state whether foreign country access to EPA GILS data is on the rise, as the rise in Table 6 demonstrates.

Perhaps one of the more innovative log analysis techniques developed for this study is that of path analysis. Path analysis enables a Web server administrator to determine a user's path and actions through a server for any given session. Table 7 demonstrates the possibility of such an analysis technique (the full IP address of the user was removed to protect the identity of that user). As Table 7 demonstrates, user xxx.olin.com first accessed the EPA server at the server's home page at 7:44AM. The user accessed a variety of EPA pages and generated several hits. At 7:46AM, the user entered GILS Earth100 directory, accessed a variety of files, and then exited the Earth100 directory at 7:54AM. The user remained logged onto the server, but did not perform any additional actions until 12:38PM. At that time, the user browsed and performed a variety of searches before logging off the server at 12:43PM. In all, the user remained logged onto the EPA server for 5 hours and 59 minutes.

In summary, it is possible to demonstrate the following through the EPA access log data:

- The total number of daily and weekly accesses and hits;

- The average daily accesses and hits;

- The percentage of accesses and hits generated by EPA's GILS, as measured through the Earth100 directory;

- The percentage of accesses and hits that the EPA server in general and EPA's GILS in particular (as measured through the Earth100 directory) experienced from the US and foreign countries; and

- The specific path a user takes through a server per session.

Together, these statistics provide an understanding of the overall use of the EPA Web server, as well as particular information resources (e.g., GILS) provided by the server.

Table 1. EPA Server Number of Hits/Accesses for February 2, 1997 to February 8, 1997.

Date	Hits	Accesses
2-Feb	131,094	42,461
3-Feb	296,687	100,437
4-Feb	307,714	104,321
5-Feb	301,650	101,530
6-Feb	278,000	93,348
7-Feb	260,330	85,789
8-Feb	113,121	41,440
Week one totals	**1,688,596**	**569,326**
Week one averages	**241,228**	**81,332**

Table 2. EPA Server Number of Hits/Accesses for February 9, 1997 to February 15, 1997.

Date	Hits	Accesses
9-Feb	131,015	52,055
10-Feb	78,174	26,358
11-Feb	267,864	96,631
12-Feb	322,080	123,782
13-Feb	306,930	116,504
14-Feb	280,737	109,583
15-Feb	109,327	39,863
Week two totals	**1,496,127**	**564,776**
Week two averages	**213,732**	**80,682**

Table 3. EPA Server Earth100 Directory Number of Hits/Accesses for February 2, 1997 to February 8, 1997.

Date	Hits	% of EPA Server	Accesses	% of EPA Server
2-Feb	292	.22%	232	.55%
3-Feb	581	.20%	449	.45%
4-Feb	684	.22%	550	.53%
5-Feb	639	.21%	502	.49%
6-Feb	642	.23%	498	.53%
7-Feb	589	.23%	407	.47%
8-Feb	417	.37%	339	.82%
Week one totals	**3,844**	**.23%**	**2,977**	**.52%**
Week one averages	**549**	**.23%**	**425**	**.52%**

Table 4. EPA Server Earth100 Directory Number of Hits/Accesses for February 9, 1997 to February 15, 1997.

Date	Hits	% of EPA Server	Accesses	% of EPA Server
9-Feb	462	.35%	331	.64%
10-Feb	225	.29%	163	.62%
11-Feb	902	.34%	625	.65%
12-Feb	1,097	.34%	771	.62%
13-Feb	853	.28%	589	.51%
14-Feb	792	.28%	603	.55%
15-Feb	493	.44%	369	.93%
Week two totals	**4,824**	**.32%**	**3,451**	**.61%**
Week two averages	**689**	**.32%**	**493**	**.61%**

Table 5. EPA Server and Earth100 Directory Percentage of Accesses From Country of Origin for February 2, 1997 to February 8, 1997.

	EPA		GILS (Earth100)	
Date	**US Accesses**	**Outside US**	**US Accesses**	**Outside US**
2-Feb	73.09%	26.91%	76.34%	23.66%
3-Feb	71.07%	28.93%	62.96%	37.04%
4-Feb	70.98%	29.02%	68.25%	31.75%
5-Feb	70.96%	29.04%	70.70%	29.30%
6-Feb	71.63%	28.37%	70.86%	29.14%
7-Feb	71.50%	28.50%	69.91%	30.09%
8-Feb	78.11%	21.89%	82.22%	17.78%
Week one averages	**72.48%**	**27.52%**	**71.61%**	**28.39%**

Table 6. EPA Server and Earth100 Directory Percentage of Accesses From Country of Origin for February 9, 1997 to February 15, 1997.

	EPA		GILS (Earth100)	
Date	**US Accesses**	**Outside US**	**US Accesses**	**Outside US**
9-Feb	79.51%	20.49%	80.05%	19.95%
10-Feb	64.51%	35.49%	46.20%	53.80%
11-Feb	73.97%	26.03%	70.59%	29.41%
12-Feb	77.24%	22.76%	74.27%	25.73%
13-Feb	76.75%	23.25%	73.46%	26.54%
14-Feb	76.14%	23.86%	61.96%	38.04%
15-Feb	71.21%	28.79%	48.04%	51.96%
Week two averages	**74.19%**	**25.81%**	**64.94%**	**35.06%**

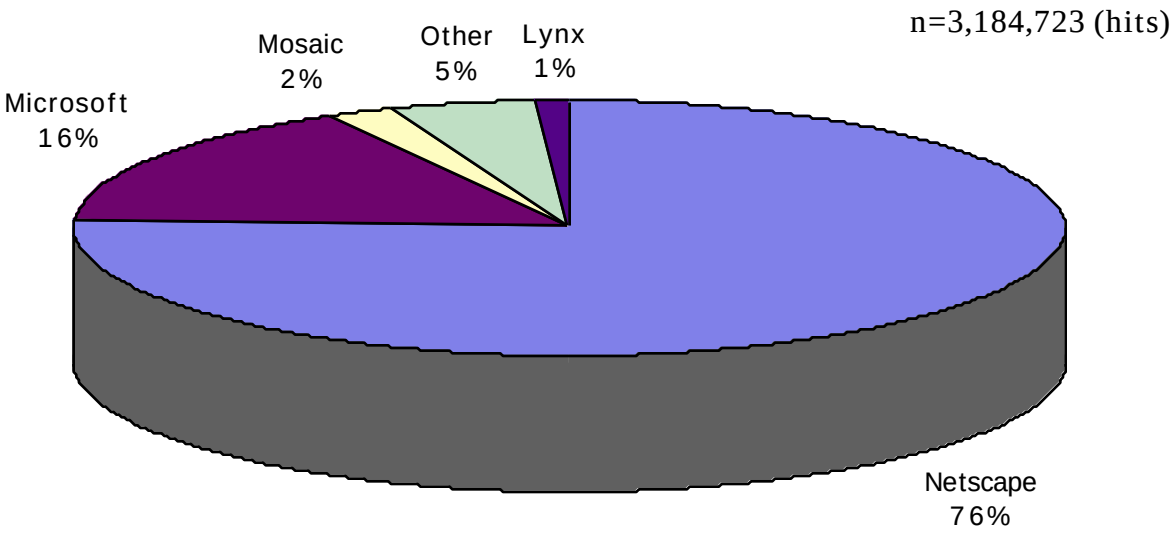
Table 7. EPA GILS Path Analysis for User xxx.olin.com for February 12, 1997.

Path and Time
/ 07:44:17
/epahome/images/2title1n.gif 07:44:21
/epahome/images/browse.gif 07:44:21
/epahome/images/newmenu.gif 07:44:21
/epahome/images/2message.gif 07:44:22
/epahome/images/newmenu.map?436,182 07:44:56
/epahome/finding.html 07:44:58
/epahome/images/epahr1.gif 07:45:04
/epahome/images/2searc1n.gif 07:45:04
/Access/index.html 07:45:38
/Access/images/epaseal.gif 07:45:41
/icons/construction.gif 07:45:42
/earth100/ 07:46:43
/cgi-bin/odometer.gif?/gils.html&width=6&.gif 07:46:47
/earth100/fish.gif 07:46:48
/oar/images/exit.gif 07:46:48
/earth100/browse.html 07:47:17
/earth100/browse/index.html 07:47:28
/icons/epabar.gif 07:47:31
/earth100/browse/C.html 07:47:52
/earth100/records/a00108.html 07:48:29
/earth100/records/a00195.html 07:51:06
/earth100/browse/L.html 07:52:52
/earth100/browse/E.html 07:53:56
/earth100/browse/I.html 07:54:19
/ 12:38:47
/epahome/images/2message.gif 12:38:57
/epahome/images/newmenu.gif 12:38:57
/epahome/images/browse.gif 12:38:57
/epahome/images/2title1n.gif 12:38:57
/epahome/images/newmenu.map?310,37 12:39:27
/epahome/images/browse.map?391,16 12:39:46
/epahome/search.html 12:39:55
/epahome/search.html 12:39:56
/epahome/images/2title1n.gif 12:39:58

Table 7. EPA GILS Path Analysis for User xxx.olin.com for February 12, 1997.

Path and Time
/oar/images/exit.gif 12:39:58
/epahome/images/2searc1n.gif 12:39:58
/epahome/images/epahr1.gif 12:39:59
/cgi-bin/waisgateII 12:40:15
/epahome/images/epa.gif 12:40:25
/cgi-bin/waisgateII?WAISdocID=6921415498+30+0+0&WAISaction=retrieve 12:41:25
/epahome/mapping.htm 12:41:35
/epahome/images/2searc1n.gif 12:41:45
/epahome/images/epahr1.gif 12:41:45
/cgi-bin/waisgateII?WAISdocID=6921415498+30+0+0&WAISaction=retrieve 12:42:07
/epahome/images/2title1n.gif 12:42:15
/epahome/images/2searc1n.gif 12:42:18
/epahome/images/epahr1.gif 12:42:19
/cgi-bin/waisgateII?WAISdocID=6921415498+0+0+0&WAISaction=retrieve 12:42:32
/cgi-bin/waisgateII?WAISdocID=6921415498+0+0+0&WAISaction=retrieve 12:42:51
/icons/epabar.gif 12:42:58
/epahome/images/2title1n.gif 12:43:23
Arrived: 07:44:17 Left: 12:43:23
Total time: 5:59:05

Figure 1. EPA Agent Log for February 2, 1997 to Febuary 15, 1997 by Type of Browser.



4.2. Agent Log

From the agent log, it is possible to determine the type of browser that users use when accessing a particular site. As Figure 1 demonstrates, Netscape is the browser of choice for a vast majority of EPA server users with 76%, followed by Microsoft Internet Explorer with 16%, Other (e.g., AOL browser, GNN)[1] with 5%, Mosaic with 2%, and Lynx with 1%.

As Figure 2 indicates, the agent log can be further analyzed to determine what version of a particular browser used to access a site. A vast majority of users -- 70.49% -- use a version of Netscape that is 2.0 or later. Thus, most all users can access forms-based Web data (a feature incorporated into later versions of Netscape). Only 41.06% of Netscape users, however, can access Java-based Web data (a feature incorporated into Netscape 3.0).

Figure 3 shows that a vast majority of users accessing the EPA Web site use a PC platform with 47% using Windows, 35% using Windows95, and 2% using Windows NT. Mac users account for only 9% of the EPA server traffic.

In summary, the agent log data provides the following data:

- The type of browser used to access a Web site;
- The version of a browser used to access a Web site; and
- The operating system of the computer used to access a Web site.

These data are particularly important as they indicate to the Web server administrators the access and display capabilities of users.

[1]To recognize all the various browsers, the PERL scripts need to specifically look for each browser. The scripts used for this analysis counted the major existing browsers.

Figure 2. EPA Agent Log for February 2, 1997 to Febuary 15, 1997 by Version of Netscape Browser.

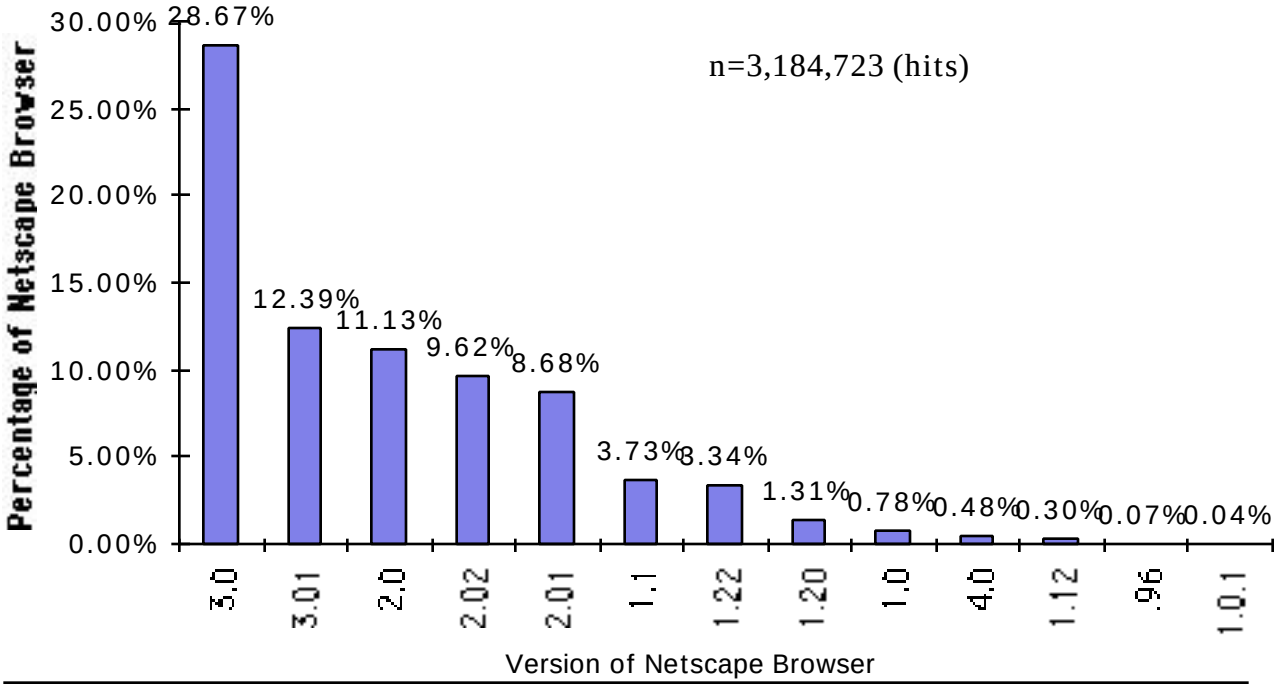
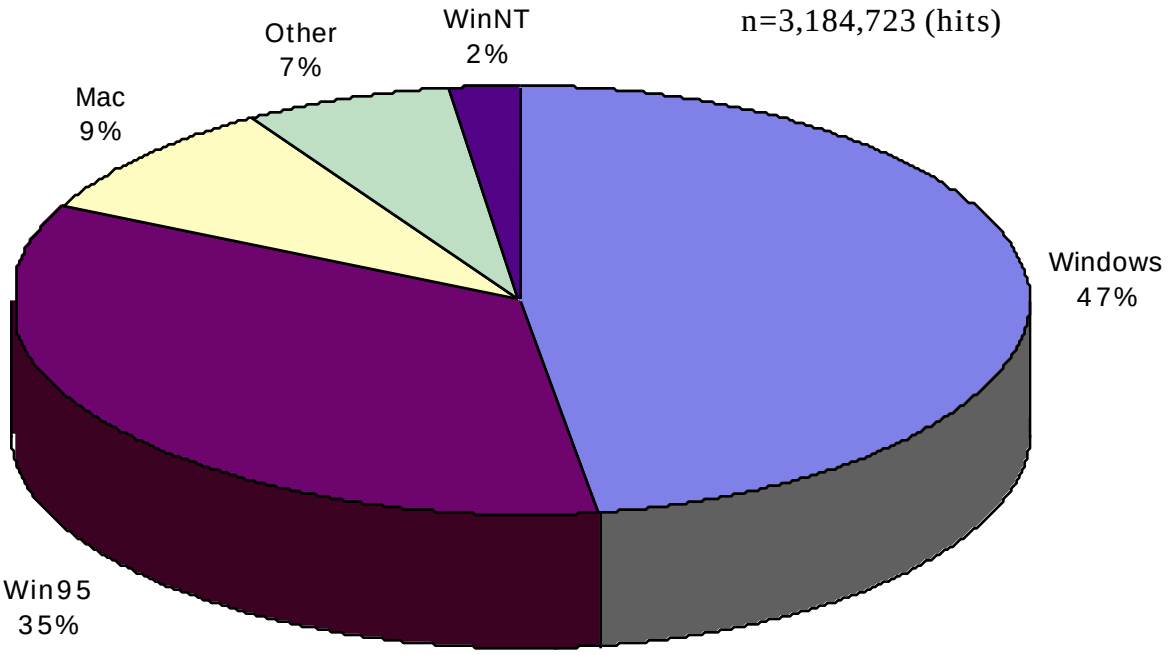


Figure 3. EPA Agent Log for February 2, 1997 to Febuary 15, 1997 by Type of Operating System.



4.3. Referer Log

The referer log indicates the Web site from which a user enters the Web server of interest. For example, if a user conducts a search using the AltaVista search engine and finds a retrieved search item of interest, the receiving Web site will show that AltaVista "refered" the user to their site. The referer log will also indicate the number of erroneous referrals (e.g., problematic links).

Tables 8 and 9 show that the EPA server received an average of 7,900 daily unique referrals (i.e., unique IP addresses) and an average of 11,500 total referrals (including multiple referrals from the same IP addresses). As indicated in Tables 8 and 9, the week of February 9, 1997 to February 15, 1997 had considerably more referral errors -- 29,937 as compared to 19,967 for the week of February 2, 1997 through February 8, 1997. It is not possible to determine precisely the reason for this increase in errors. Tables 8 and 9 also indicate that the GILS directory receives relatively few referrals as compared to the EPA server in general (a daily average of 5 referrals). It is interesting to note, however, that most of the GILS referrals are from unique IP addresses, indicating that users access the GILS directory from different sources each time.

The EPA server receives so many referrals that it is not feasible to identify all referring sources. Based on the analysis of the referer log, however, it was possible to identify the four most frequently referring sites to the EPA server (see Tables 10 and 11). The most frequently referring sites to the EPA server are the Yahoo-Society site with an average daily referral rate of 68 and 77 respectively, the Yahoo-Government site with an average daily referral rate of 60 and 66 respectively, the Web Directory with an average daily referral rate of 46 and 55 respectively, and the Web Crawler with an average daily referral rate of 30 and 39 respectively.

In summary, it is possible to generate the following data from the referer log:

- The total and unique number of referring IP addresses; and
- The most frequently referring sites.

This type of data enables Web server administrators to determine who their most frequent referring sites are as well as the overall number of referrals. Such data is particularly useful for Web server administrators as changes to document links are made on a server. With the assistance of the referer log data, server administrators of a site can contact the most frequently referring sites to make changes in their links so as to avoid user-encountered error messages.

Table 8. EPA Server and Earth100 Directory Referrals for February 2, 1997 to February 8, 1997.

	EPA			GILS (Earth100)	
Date	**Total Referrals**	**Unique Referrals**	**Total Errors**	**Total Referrals**	**Unique Referrals**
2-Feb	7,318	5,462	890	2	2
3-Feb	13,946	9,076	1,638	3	2
4-Feb	15,119	9,786	2,440	8	8
5-Feb	14,304	9,199	4,708	3	3
6-Feb	13,506	8,889	3,931	24	8
7-Feb	12,320	8,330	3,980	27	8
8-Feb	6,763	5,151	2,380	1	1
Week one totals	**83,276**	**55,893**	**19,967**	**68**	**32**
Week one averages	**11,897**	**7,985**	**2,852**	**10**	**5**

Table 9. EPA Server and Earth100 Directory Referrals for February 9, 1997 to February 15, 1997.

	EPA			GILS (Earth100)	
Date	**Total Referrals**	**Unique Referrals**	**Total Errors**	**Total Referrals**	**Unique Referrals**
9-Feb	7,570	5,644	2,657	2	2
10-Feb	13,777	9,058	4,396	7	7
11-Feb	13,978	9,370	5,045	9	8
12-Feb	13,941	9,484	5,089	3	3
13-Feb	14,022	9,262	5,707	7	7
14-Feb	11,647	7,885	4,513	5	5
15-Feb	6,061	4,628	2,530	5	5
Week two totals	**80,996**	**55,331**	**29,937**	**38**	**37**
Week two averages	**11,571**	**7,904**	**4,277**	**5**	**5**

Table 10. EPA Server Top Referring Sites for February 2, 1997 to February 8, 1997.

Date	Web Directory	Yahoo -Society	Yahoo-Government	Web Crawler
2-Feb	46	43	36	29
3-Feb	58	76	82	55
4-Feb	65	91	83	36
5-Feb	59	82	92	41
6-Feb	80	80	84	45
7-Feb	52	61	65	41
8-Feb	27	44	22	25
Week one totals	**387**	**477**	**464**	**272**
Week one averages	**55**	**68**	**66**	**39**

Table 11. EPA Server Top Referring Sites for February 9, 1997 to February 15, 1997.

Date	Web Directory	Yahoo -Society	Yahoo-Government	Web Crawler
9-Feb	36	60	37	28
10-Feb	75	89	91	46
11-Feb	53	82	79	28
12-Feb	49	89	68	24
13-Feb	33	101	70	35
14-Feb	42	67	54	32
15-Feb	33	48	20	19
Week two totals	**321**	**536**	**419**	**212**
Week two averages	**46**	**77**	**60**	**30**

4.4. Error Log

The error log provides data on the errors (e.g., dead links, aborted file downloads) that users either encounter or initiate. Due to the large volume of traffic that the EPA Web server generates, it is not feasible to present all the error log data. As such, selected error data is presented in this section.

Tables 12 and 13 show that six files are consistently aborted by users. These include the cgi-bin/waisgateII (an average daily download abort rate of approximately 200), cgi-bin/waisgate (an average daily download abort rate of approximately 160), /oar/oarmap.gif (an average daily download abort rate of approximately 50), /icons/nceri2.gif (an average daily download abort rate of approximately 40), epahome/404.html (an average daily download abort rate of approximately 35), and OW/images/feb_ad.gif (an average daily download abort rate of approximately 32).

Tables 14 and 15 demonstrate both the total number of download aborts as well as the most consistently aborted downloaded file for the GILS Earth100 directory. The data indicate that the GILS directory has a daily average of 5 send aborts, most of which are due to the isi.zip file (see Tables 14 and 15).

In summary, the error log data provide the following data:

- The overall and directory specific number of aborted downloads; and
- The overall and directory specific most frequently aborted file downloads.

This type of data is helpful to Web administrators as it indicates which images, files, and pages require too much time to download, leading users to abort the files.

The next section presents key issues that the study team encountered in developing the PERL scripts for EPA log analysis, the formatting of the EPA log files, and the general management of EPA's log files.

Table 12. EPA Server Most Frequently Aborted Downloads for February 2, 1997 to February 8, 1997.

Date	cgi-bin/ waisgate	cgi-bin/ waisgateII	/oar/ oarmap.gif	epahome /404.html	OW/images /feb_ad.gif	/icons /nceri2.gif
2-Feb	65	42	30	34	10	16
3-Feb	248	316	82	58	44	66
4-Feb	291	314	59	60	39	78
5-Feb	266	258	88	44	42	52
6-Feb	250	216	76	33	48	49
7-Feb	179	173	44	40	44	43
8-Feb	54	92	22	23	17	13
Week one totals	**1,353**	**1,411**	**401**	**292**	**244**	**317**
Week one averages	**193**	**202**	**57**	**42**	**35**	**45**

Table 13. EPA Server Most Frequently Aborted Downloads for February 9, 1997 to February 15, 1997.

Date	cgi-bin/ waisgate	cgi-bin/ waisgateII	/oar/ oarmap.gif	epahome /404.html	OW/images /feb_ad.gif	/icons /nceri2.gif
9-Feb	43	110	24	13	16	9
10-Feb	237	264	50	28	31	38
11-Feb	245	310	74	46	39	54
12-Feb	225	257	64	48	59	51
13-Feb	170	234	58	42	33	37
14-Feb	180	207	55	61	38	68
15-Feb	46	102	18	35	18	10
Week two totals	**1,100**	**1,382**	**325**	**238**	**216**	**257**
Week two averages	**157**	**197**	**46**	**34**	**31**	**37**

Table 14. EPA Server Total and Most Frequently Aborted Downloads for the Earth100 Directory for February 2, 1997 to February 8, 1997.

Date	Total send aborts	Send aborts on isi.zip
2-Feb	7	6
3-Feb	6	2
4-Feb	2	1
5-Feb	5	2
6-Feb	3	3
7-Feb	9	4
8-Feb	0	0
Week one totals	**32**	**18**
Week one averages	**4.57**	**2.57**

Table 15. EPA Server Total and Most Frequently Aborted Downloads for the Earth100 Directory for February 9, 1997 to February 15, 1997.

Date	Total send aborts	Send aborts on isi.zip
9-Feb	4	2
10-Feb	13	3
11-Feb	7	4
12-Feb	2	0
13-Feb	4	1
14-Feb	6	1
15-Feb	5	4
Week two totals	**41**	**15**
Week two averages	**5.86**	**2.14**

5.0. KEY DATA AND LOG FILE ANALYSIS ISSUES AND RECOMMENDATIONS

Throughout the log analysis process, the study team encountered a number of problems and issues that affected its ability to develop log analysis script files and perform certain types of log analysis. Below is a list of key issues encountered and recommendations to resolve such problems:

- *Transfer of files*. The study team had no guarantee that the files it received were the complete data set. There is a need to implement a procedure whereby EPA would post the file size of the log files directly from the server and the study team could verify this against the downloading files. An example of this is for the February 10, 1997, for which there was clearly data missing from the access log.

- *Storage space.* Storing just two weeks of log files from the EPA, as well as the PERL scripts and the resulting files took up nearly a gigabit of hard drive space. If the development and analysis of EPA's log files were to continue, the study team would need to dedicate a machine with adequate hard drive space to maintain the files. Moreover, as the study team suffered a server crash during this study, it is also necessary to have a back up server or tape backup of the script and log data files.

- *Enhancing the access log scripts*. Given time and resource constraints, a majority of the analysis for this project was done using Microsoft Excel. It is possible to incorporate some of this analysis into a re-write of the PERL scripts. For example, the percentages of U.S. hits versus outside hits in the access log were added manually. This calculation would be relatively simple to have the PERL scripts create.

- *Accommodating different log file formats*. A portion of the PERL scripts used for this project were originally written and tested with Syracuse University-generated log files. The study team found, however, that the EPA's Web server used substantially different file formats. For example, the original PERL scripts did not count files with .txt extensions or cgi-bin files as accesses, a feature required for the accurate analysis of the EPA server files. The study team re-wrote the scripts prior to the final analysis presented in this study. In the future, though, it would be necessary to have a team of people responsible for assessing the log file format and composition of the Web server prior to using the scripts to ensure that no file formats and/or other data were missing.

- *Awareness of script and counting errors*. The study team encountered two main errors in their PERL script development: (1) double counting hits, and (2) erroneous counting of EPA main page accesses. The study team initially developed PERL scripts to count the number of server hits as well as the percentage of accesses directly to the EPA server's main page. However, the scripts erroneously double counted some hits and could not accurately track the main page accesses. The hit counting errors were corrected, but the percentage of main page access could not be. The latter data were not reported in this study.

- *Separating in-house from external referrals*. There are a large number of EPA-based hosts that access the EPA server (e.g., www.epa.gov, ftp.epa.gov, earth1.epa.gov). It is, therefore, difficult to completely remove EPA-based accesses and hits from the referral logs. While the study team attempted to remove such internal referrals to get a more accurate picture of the server's non-EPA use, the study team is certain that they were unable to adjust for all the EPA domain names that might exist.

- *Separating various search engines*. The study team underestimated the number of daily referrals that the EPA server received. From analyzing the log analysis data, it is clear that a number of the referrals come from search engines. The PERL scripts were not written to extract this information. Future development of the scripts can help the study to determine not only what percentage of referrals come from search engines, but what search engines users tend to use and what search terms users enter.

- *Specific path and error analysis*. At present, the PERL scripts can only analyze a specific directory from the access log (e.g, the GILS [Earth100] directory). Incorporating error log information into the referer log analysis, however, required a custom shell script. More work is needed to fully incorporate error and referer log file data.

- *Extended log format*. There are multiple types of log file formats. The EPA Web server currently generates log files in the common log format. The extended log file format, however, allows all log information to be collected into one log file. Although this would mean the study team's PERL scripts

would require a complete revision, it would be possible to collect more information about specific visitors through this file format.

The above issues provide an insight into the key problems that the study team encountered and attempted to resolve while performing the log analysis of the EPA Web server log files. The problems illustrate the newness of log file analysis, the lack of consistency of log file formats, and the need to develop additional means of analyzing Web server log files.

The next section presents key issues in the collection, use, and interpretation of log file data.

6.0. KEY LOG FILE INTERPRETATION AND MANAGEMENT ISSUES

There are several key issues that Web service providers should consider when using log files as indicators of digital service output measures. These include:

- *Interpreting and considering the log files as one component of a larger assessment activity for networked services*. While log files can provide Web administrators and others with critical server-related data, log files do not reflect user-based impact and outcome measures. Log files, therefore, combine both user and technical perspectives on Web services.

- *Understanding what, exactly, the data reflect*. The distinction between "hits" (downloads on an html page) and accesses (a downloaded html page) is critical. Software that counts only "hits" will not reflect the true nature of the site's use. In addition, neither "hits" nor accesses translate directly into distinct users. Many Internet service providers, such as America OnLine, use "proxy" servers. Because of this, the Access Log will not accurately trace the number of users but, rather, reflect the number of accesses/"hits" made by the referring server.

    - A related issue is understanding the context of the server and presenting the data within that context. For example, this study concentrated on the use of the GILS Web server in the context of the EPA Web server. Readers cannot not, however, interpret the Web-based GILS record use as scant, moderate, or high without knowing the usage of each EPA server component. This study did not seek to provide that context.

- *Knowing what data to count*. Each Web server has different file naming conventions and methods of organization. For example, the EPA server used such file extensions as .txt to designate Web pages (as opposed to the more commonly used .htm or .html extensions). In order to accurately reflect the page accesses, the study team re-wrote the PERL scripts to count .txt extension files as accesses rather than hits. It is not clear to the study team, however, if these naming conventions hold throughout the entire server. Therefore, some accesses may actually be represented as hits in this study.

    - A related, and important issue, is that of internal versus external EPA server use. EPA has several IP domains that access the EPA server on a daily basis. The issue is the extent to which some of those accesses and/or hits are due to public requests for information. There is no current way, as of today, to gather such data. In the future, however, it may be worth identifying, isolating, and analyzing server use by a selection of EPA domain addresses that serve as public information offices, for example, to gain a greater sense of the EPA server's public service provision activities.

- *Selecting and/or developing appropriate analysis software*. Web server administrators need to plan for the analysis of Web server log files. The types of information about Web server use desired by those running the server should drive the selection and/or development process of log analysis software. Web administrators should not retrofit their log file analysis to the capabilities of the software.

- *Obtaining the cooperation of server administrators and Internet service providers*. Not all networked information providers run their own servers or have direct control over the Web server on which the Web-based services reside. As such, it is important to gain the cooperation of those individuals and/or entities that have direct control over the log files. The lack of such cooperation will have a negative impact on the ability to attain Web server usage data.

- *Preserving the privacy and confidentiality of server users*. In some cases, it is possible to trace directly back to a user, depending on the method of access a user has to a Web site. Web service providers need to develop policies as to how such data, if at all, will be used. This issue is particularly troublesome for public sector organizations, as such capabilities may violate privacy laws.

- *Educating server administrators and decision makers as to the benefits of log file analysis*. Log file analysis is just beginning to gain popularity. Server administrators and decision makers need to understand the types of data that log files can generate, the application of that data in an organizational setting, and the incorporation of such data into management activities.

- *Managing the log analysis process*. Gaining access to and analyzing Web server log files requires planning and coordination. To engage in log file analysis activities, there needs to be a delegation of responsibility for making the files available (on-site or remotely), performing the analysis (on-site or remotely), interpreting the analyzed data, and reporting the findings. Moreover, such analysis needs to be performed and reported on an ongoing and regular basis.

- *Presenting Web log statistics effectively on the Web itself*. Two issues require resolution: (1) the presentation of Web usage statistics on the Web; and (2) the means through which to display such statistics. Increasingly, users want to review Web server statistics of the sites they visit. This requires the presentation of those statistics by the site providers. Since Web usage statistics are in their infancy, however, little is known about appropriate ways in which to display such usage data and the purpose that is served in doing so.

These issues serve as a beginning point for Web server log analysis collection, presentation, and interpretation. Other issues exist, and still more will develop as Web services increase and log analysis techniques become more sophisticated.

7.0. MOVING FORWARD

Research into the analysis of Web server log files is limited. Web server administrators and decision makers are just beginning to understand the potential for systematic server usage data, and researchers are only just beginning to develop sophisticated analysis techniques. Key areas that require further exploration include the:

- *Ability to export files and analyze them in other formal statistics programs*. Current analysis techniques require specialized software and/or the development of specific analysis programs. There is a need to develop means through which log files can be imported and analyzed using off-the-shelf statistical analysis programs.

- *Understanding of log file data as user-based measures of Web services*. By performing Web log file analysis, server administrators and decision makers can begin to understand the path users take through a server, the problems users encounter during a session, and technology users use while navigating a site. Together, these are powerful data that can assist in the planning and design of Web-based services.

- *Cross-referencing log files.* This is an area of analysis that intends to cross-tabulate the various log files. For example, by cross-referencing the Error and Access Log files, one could know how many users, after receiving an error, stop surfing the site on which the error was received. To find this percentage a server administrator would use the domain name and time of the user who received an error (from the Error Log) and then look in the Access Log to see if that domain name shows up after the time of the error.

- *Creation of script files that can assess multiple types of log files*. While there are certain log file standards, not all Web server log files are exactly alike. Until such time as all log files are the same, the development of log analysis scripts will need to be able to accommodate multiple log file types so as to generate the same types of information regardless of file type.

- *Customization of script files*. Even if all Web servers generate log files that conform to certain standards, there will likely be differences in Web page and file naming conventions across servers. As such, script files will require modifications to meet the needs of specific log data.

- *Separation of internal versus external server traffic*. In order to determine the user community of the server, Web server administrators need to know who is accessing the server. Cursory analysis of the EPA log files indicates that a substantial proporation of the EPA Web server is generated by EPA IP addresses. It would be useful to know if EPA-generated Web server traffic and use differs than that of non-EPA users. Such data would assist the server administrators customize various portions of the server to more specifically mee the needs of various user groups.

- *Incorporation of log file analysis with other on-going electronic network assessment techniques*. The assessment and evaluation of electronic networks and network-based resources is increasing in scope and application (see Moen & McClure, 1997; Bertot & McClure, 1996; McClure & Lopata, 1995). Web log file analysis is a network-based assessment technique that is particularly useful when performed in conjunction with other on-going evaluation activities.

There is a need to resolve, minimally, these issues and move the ability to perform log file analysis forward. Log file data can provide user-based measures of Web-based resources if performed on a regular basis, incorporated into other electronic network assessment activities, and interpreted correctly.

This study presents a beginning point for Web log file assessment techniques. Researchers, server administrators, and decision makers are just now starting to understand the potential for Web log file analysis as part of a larger user-based measure of electronic resources. As the Federal government increases its provision of Web-based services for its citizens, agencies will need to develop, implement, and maintain an on-going assessment of Web-based activities through the analysis of Web server log files.

REFERENCES

Bertot, John Carlo & McClure, Charles R. (1996). *Sailor Assessment Final Report: Findings and Future Sailor Development.* Baltimore, MD: Division of Library Development and Services.

Christian, Eliot J. (1996, December). GILS: What is it? Where is it going? *D-Lib Magazine*. <http://www.ukoln.ac.uk/dlib/dlib/december96/12christian.html>.

Christian, Eliot J. (1994, May). *The Government information locator service (GILS): Report to the information infrastructure task force*. < http://www.usgs.gov/gils/ gilsdoc.html>.

McClure, Charles R., and Lopata, Cynthia. (1996). *Assessing the Academic Networked Environment: Strategies and Options*. Washington, D.C.: Coalition for Networked Information.

McClure, Charles R., Moen, William and Ryan, Joe. (1992). Design for an Internet-Based Government-Wide Information Locator System. *Electronic Networking: Research, Applications, and Policy, 2*, 6-37.

Moen, William E. & McClure, Charles R. (1997, January 14). Multi-Method Approach for Evaluating Complex Networked Information Services: Report from an Evaluation Study of the Government Information Locator Service (GILS). Technical Paper Submission: ACM Digital Libraries '97. [Paper Submitted].

Moen, William E. & McClure, Charles R. (1996, August). *Technical Proposal: An Evaluation of the Federal Government's Implementation of the Government Information Locator Service (GILS).* Denton, TX: University of North Texas, School of Library and Information Sciences. <http://www-lan.unt.edu/slis/research/gilseval/ gilseval.htm>.

Moen, William E. & McClure, Charles R. (1994). *The Government Information Locator Service (GILS): Expanding Research and Development on the ANSI/NISO Z39.50 Information Retrieval Standard, Final Report*. Prepared for the United States Geological Survey and the Interagency Working Group on Data Management for Global Change, Washington, DC [USGS, Cooperative Agreement No., 1434-93-A-1182]. Bethesda, MD: NISO.

National Institute of Standards and Technology, Computer Systems Laboratory. (1994, December 7). *Federal Information Processing Standards Publications 192, Application Profile for the Government Information Locator Service (GILS).* Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology. <http://www.dtic.dla.mil/gils/documents/naradoc/fip192.html>.

Network Wizards. (1997). *Internet Domain Survey, January 1997.* <http://www.nw.com/>.

Noonan, Dana. (1996). Making Sense of Web Usage Statistics. *The PIPER Letter: Users and Usage*. <http://www.piperinfo.com/piper/pl01/usage.html>.

Novak and Hoffman. (1996). New Metrics for New Media: Toward the Development of Web Measurement Standards. <http://www2000.ogsm.vanderbilt.edu/novak/web.standards/webstand.html>.

Rubin, Jeff. (1996). Log Analysis - A Brief Overview. <http://headcase.syr.edu/text/logs.html>.

Stout, Rick. (1997). *Web Site Stats: Tracking Hits and Analyzing Traffic*. Osborne McGraw Hill: Berkley, CA.